Achieving Automated Narrative Text Interpretation Using Phrases in the Electronic Medical Record

Shawn N. Murphy M.D., Ph.D. and G. Octo Barnett M.D.
Laboratory of Computer Science
Massachusetts General Hospital, Boston, MA.

Stereotypic phrases are used by clinicians throughout the medical record, as seen in an analysis of our COSTAR medical record database. These phrases are often associated with an underling semantic concept; for example the phrase CLEAR LUNGS may be linked with the concept "normal lung exam" for a particular physician. Formalizing these associations with concepts from the UMLS using the MEDPhrase application allowed us to automate interpretation of narrative text within our electronic medical record.

INTRODUCTION

The creation of a patient note is an essential and time consuming task for the clinician. In general, clinicians would rather create a note using narrative text which is familiar, than narrative text which is unfamiliar and seems manufactured by a computer. In part, this is because clinicians develop habits of expression over years of repetition.

During the creation of a typical note, the clinician will often use well-practiced phrases which change little from one patient to the next. Many such phrases are in common use by clinicians. For example, the phrases LUNGS CLEAR or CLEAR LUNGS were found 38,550 times in 117,109 chest exams from our COSTAR database. The average length of the chest exam was 2.3 words, and many entries consisted of just those two words. This consistent use of phrases was found throughout many sections of the medical record.

The phrases used by a clinician can often be mapped to a concept that the clinician is trying to express. In the case of the phrase CLEAR LUNGS, the intent of the clinician is usually to express the concept of a normal chest exam.

There is a need in the field of medical informatics for a clinical note to expresses unambiguous concepts. Previous applications have been developed that created narrative text out of data elements selected through the user interface.[1,2] This may assure that concepts expressed in the text are made up of data elements recognized in the database. The cost of this assurance is the time-consuming task of construction of every phrase from the most basic data elements, even for commonly used phrases.

The use of natural language processing offers the promise of saving time for the clinician while allowing concepts to be extracted from narrative text.[3,4,5] Difficulties with these methods are both the extraordinary complexity of language, and its occasionally idiosyncratic use by clinicians. The same phrase may also have different meanings to different clinicians. Furthermore, the context of the phrase may change the concept expressed by similar phrases.

A less ambitious form of natural language processing can occur by encouraging clinicians to use those phrases they are already using in their notes, but link the phrases to a controlled vocabulary, and restrict the context in which the phrases can be used. The clinician can develop a phrase library consisting of phrases that are commonly used in his or her patient notes. The use of a phrase library to construct a patient note has been previously presented.[6] We propose to link the stereotyped phrases to a controlled vocabulary and restrict the context in which the phrases may be used so that a note containing narrative text can be used to express concepts unambiguously.

THE APPLICATION

In order to assist with the creation of clinical notes using clinician specific phrases linked to a controlled vocabulary, we created a stand-alone Microsoft Windows application called MEDPhrase. The visual presentation of MEDPhrase is shown in Figure 1, and consists of a memo-type pad that may be located on the screen next to a text editor of an electronic medical record (EMR) system. Within the memo-type pad appears a phrase list, or rather, the labels of phrases in the phrase list. Labels are used for phrases when the phrases themselves are too large to be shown unambiguously on a single, short line. A phrase is transferred to a text editor of the medical record by a mouse-click on the label of a desired phrase, or by using a series of key strokes to identify and transfer the desired phrase.

0195—4210/96/$5.00 © 1996 AMIA, Inc.

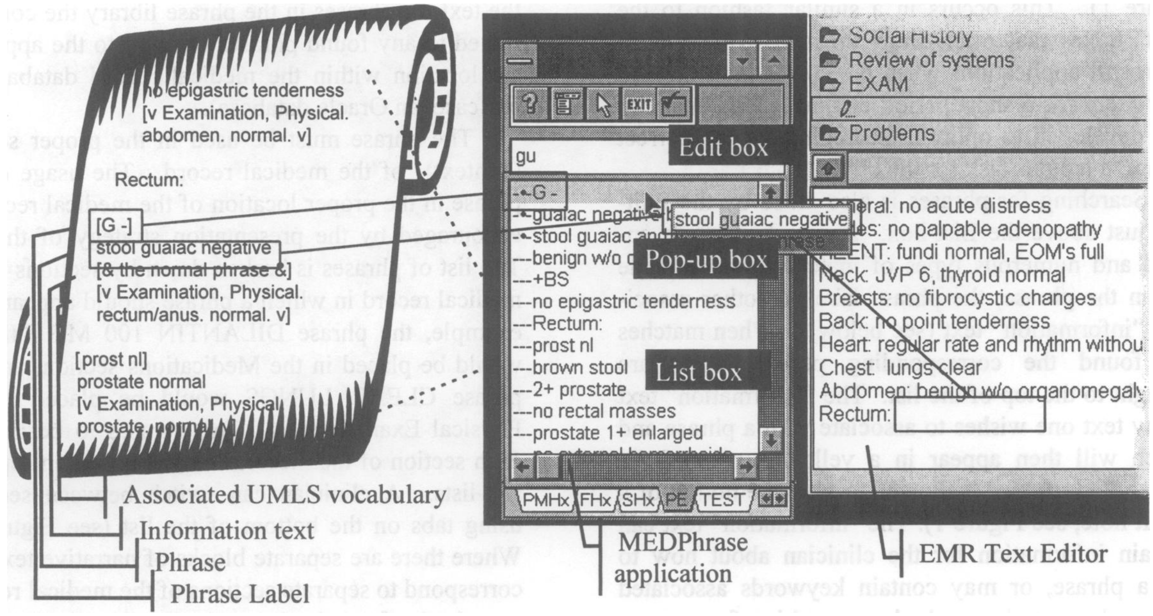

Figure 1. Transferring phrases to the EMR Text Editor. Part of the phrase list used in this example is shown on the left in our standard phrase list markup language. Entries are separated by blank lines. The phrases themselves have no mark up, the phrase title is placed between "[ ]," the "information" text is placed between "[& &]," and the linked UMLS vocabulary is placed between "[v v]." Phrases without linked UMLS vocabulary do not store concepts in the EMR database, but may function as an aid to the clinician in composing a note. In this example, the phrase "stool guaiac negative" is about to be transferred to the EMR text editor when the cursor is clicked, to appear in the last of the daisy-chain linked boxes (going left to right). The daisy-chain linked boxes illustrate the flow from the ASCII list to the MEDPhrase application and finally to the EMR text editor. When the text "stool guaiac negative" is written to the EMR text editor, the linked UMLS vocabulary *Examination, Physical - rectum/anus - normal* will be placed in a temporary database. All the text is read back from the EMR text editor by MEDPhrase when this section of the EMR is closed; and if the phrase "stool guaiac negative" is found to be unaltered the linked UMLS vocabulary *Examination, Physical - rectum/anus - normal* will be sent and stored in the EMR database of the active patient.

Each phrase in the phrase list is attached to data elements in a controlled vocabulary. This linkage is defined for each clinician to represent the meaning intended by that user. The same phrase used by one clinician may not carry the same meaning to another clinician. For example, the phrase CLEAR LUNGS may have been attached by one clinician to the UMLS vocabulary "Examination, Physical - chest - normal." However, the phrase CLEAR LUNGS may be linked to the UMLS vocabulary "Examination, Physical - lung - Abnormal chest sounds - absent" for that clinician who does not imply the broader meaning "Examination, Physical - chest - normal" when using the same phrase. Indeed, it is this ambiguity of phrases used by different clinicians which makes natural language interpretation so difficult.[3] The cost of making the vocabulary link dependent on the clinician is the significant amount of time an encoder must take to manually link each phrase to a controlled vocabulary. It also takes the clinicians time to specify their meaning for each phrase to the encoders. No automated procedure has been developed to allow the clinician to link their own phrases. However, an interface such as PEN-Ivory[2], which aids the construction of concepts from the UMLS vocabulary, could be adapted for this purpose.

The MEDPhrase application was coded in Object Pascal using a product that has preprogrammed visual objects which encapsulate much of the user interface to Microsoft Windows. Most of the user interface consists of standard graphical user interface components which behave consistent with other Windows-based applications. The behavior of the Pop-up box was extended such that when the cursor passes over a phrase label in the List-box the entire phrase appears in its own window near the label (see

Figure 1). This occurs in a similar fashion to the 'hint' boxes that often show up next to buttons in Microsoft applications when the cursor is positioned over them for a short period of time. This behavior was developed to optimize searching for the correct phrase, a tedious task for the clinician.

Searching for phrases is also aided by the Edit-box just above the List-box. A search string is entered and numerous types of matches can be made within the phrase, the phrase label, or other associated "information" text (see below). When matches are found the corresponding phrase labels are brought to the top of the list. The "information" text is any text one wishes to associate with a phrase and which will then appear in a yellowed part of the phrase Pop-up box (yellowed to give the notion of a PostIt note, see Figure 1). The "information" text can contain information for the clinician about how to use a phrase, or may contain keywords associated with a phrase to be used when searching for a group of phrases.

MEDPhrase writes text to a window of the EMR by generating artificial keystrokes within the Windows messaging system. The text of the phrase appears at the cursor location in the window of the EMR as though it was typed on the keyboard. MEDPhrase can synchronize staying on top of other windows using Windows messages generated by the electronic medical record system. These messages can also control what sub-list is currently displayed, which depends upon what section of the medical record is currently being utilized.

When a phrase is written to the text editor of the EMR, MEDPhrase will place the controlled vocabulary attached to the phrase in a temporary database. The controlled vocabulary attached to the phrase is not transferred to the EMR database immediately, because the phrase is transferred to a text editor that allows editing of the text. There must be an assurance that the phrase was not altered after its transfer since it is the unaltered phrase that is mapped to the controlled vocabulary. Therefore, the transferred text must be checked to see that it contains the unaltered phrase. The MEDPhrase application waits for the task terminate message to be sent to the Windows operating system, and then grabs the text from the window associated with the task by using Dynamic Data Exchange (DDE) and processes the text. DDE is a method supported by Microsoft Windows to transfer text between applications. Currently MEDPhrase only checks for unaltered phrases within the text. This behavior could be extended with more elaborate natural language technology to accept minimally altered phrases. Following the review of the text for phrases in the phrase library the concepts linked to any found phrases are sent to the appropriate location within the medical record database (in our case, an Oracle database).

The phrase must be used in the proper section (context) of the medical record. The usage of the phrase in the proper location of the medical record is encouraged by the presentation strategy of the list. The list of phrases is broken down by sections of the medical record in which a phrase should appear. For example, the phrase DILANTIN 100 MG TID PO would be placed in the Medications section, and the phrase CLEAR LUNGS would be placed in the Physical Exam section. The phrases to be used in each section of the medical record appear in separate sub-lists. A clinician can switch between sections using tabs on the bottom of the list (see Figure 1). Where there are separate blocks of narrative text that correspond to separate sections of the medical record, one checks for only those phrases that should occur in that section of the EMR.

THE PHRASES

The acquisition of the phrase libraries can be done either by surveying clinicians who will use MEDPhrase and preparing a phrase library made up of their responses, or by searching a medical record database for commonly occurring phrases. We explored both approaches and found particularly useful the analysis of the content of the Physical Exam section of our COSTAR EMR database. Our intent was to pick up commonly used phrases and present them as a list to the clinician who then might select phrases that were familiar to him for his phrase list, and delete phrases he did not wish to use.

As phrases are acquired they must be attached to concepts. These concepts are expressed in a controlled vocabulary that is recognized by the database of the EMR. The phrases are not necessarily composed of atomized data elements, and therefore, are not self-encoding. Currently, each phrase needs to be manually reviewed with the clinician to determine the concept the clinician means to convey, and then this concept is composed into data elements recognized by the database.

Our analysis of the COSTAR database included 893,465 notes composed of a total of 4,008,474 words. Stop words were excluded. Each section of the physical exam (i.e. General Appearance, Skin, HEENT, etc.) was considered a separate note for the analysis. Phrases were kept in their literal form during the analysis except for simple word inversions (for example, CLEAR LUNGS and LUNGS

CLEAR) in which case a single phrase was chosen and occurrence statistics were combined. Abbreviations were kept in their literal form. When a smaller phrase was embedded within a larger phrase, but the smaller phase occurred more frequently, the words from the smaller phrase were subtracted from the larger phrase in generating usage statistics.

Table 1 presents the results of our analysis in tabular form and suggest that a generic (clinician independent) phrase library, made up of the top 9 phrases of the physical exam, would be able to generate the text for 10% of the text found in the physical exam notes. A phrase list made up of the top 131 phrases could generate the text for 30% of the text found in the physical exam notes. There were some sections of the physical exam that had an extraordinary amount of text composed of a single phrase. In the rectal exam section, 40% of the medical record was composed of the phrase STOOL GUIAC NEGATIVE or STOOL GUIAC NEG, and in the chest exam section, the phrase LUNGS CLEAR made up 28% of the text of the medical record.

Table 1. Usage of common phrases in the Physical Exam section of the COSTAR patient note database, shows the percent of the record that is composed of a number of the most commonly occurring phrases, and the total number of words constituting the sum of these phrases.

PERCENT OF RECORD	NUMBER OF PHRASES	TOTAL NUMBER OF WORDS
10%	9	18
20%	40	81
30%	131	286
40%	309	711
50%	674	1,595
100%	-	4,008,474

A major part of current development is centered about the organization of the lists into files. The files of phrases may be organized into those phrases that a clinician would associate with a specific (chief) complaint, or those phrases that a clinician would associate with a specific problem or diagnosis. A topic can be selected from a menu that then loads the proper file of phrases into MEDPhrase. A hierarchy of files can be maintained when one file directs several other files to be loaded. The organization of phrases into files is a way to keep the number of available phrases in the MEDPhrase List-box limited to those phrases that apply to a specific complaint, problem, or diagnosis. However, experience by other developers of user interfaces[2] suggests that a single list of phrases may be a better model, perhaps bringing the phrases associated with the selected topic to the top of the list if searching the list requires extensive scrolling.

Each file of phrases is organized into sections that correspond to the sections of a typical note. The sections are important because they reflect the context's in which the phrases must appear. Sections such as "Chief Complaint", "History of Present Illness", "Current Medications", etc. fits the organization of most medical records. The phrases are organized into these sections which are displayed as sub-lists. A clinician can move from one sub-list to another using the tabs at the bottom of the list (see Figure 1). There is nothing to prevent other types of sections from being created and maintained. However, to automatically display the proper sub-list when working within that section of the medical record, standardized sections are used. It is important to recognize that the organization of the phrases under sections is a way of defining the context in which the phrase must appear.

The organization of the phrases on the list is generally confined to a static ordering of phrases, either in an order specified by the user or alphabetically as the default. It may be worthwhile experimenting with a system where the placement of a phrase on the list is made dependent on which phrases are used most often, with more frequently used phrases making their way to the top of the lists. Interestingly, such strategies have proved to be more distracting than helpful in the experience of others.[7]

DISCUSSION

The MEDPhrase application was developed after recognizing the prolific use of stereotyped phrases in medical notes. Our analysis of the COSTAR database showed that significant portions of the medical record are made up of these freely dictated phrases. Because clinicians are already using this self-imposed order in their narrative text and clinical work-flow, we wished to use it as an aid in extracting concepts from the narrative text of the medical record.

The MEDPhrase application was developed to encourage the use of these highly stereotyped phrases in clinical notes. We believe it would be difficult, if not impossible, to extract the concepts embedded in these naturally occurring stereotypical phrases from the narrative text using natural language processing without a prior agreement upon specific structured phrases and linked concepts. Difficulty with any automatic extraction occurs with minor changes and nuances in the composition of the phrases. Furthermore, similar phrases may have different meanings to different authors, and even the same author may use similar phrases that have different meanings in different sections of the medical record.[3,5]

The MEDPhrase application provides a method to work with the stereotypical phrases and keep track of their meanings. The Unified Medical Language System (UMLS) was developed in order to provide a common vocabulary to facilitate explicit communication between clinicians. When the meanings of phrases are expressed in UMLS, the narrative text of the phrase becomes a link to this expression. This link is automatically lost when the phrase is modified in the text, or the phrase is used in the wrong section of the medical record (the wrong context).

There are powerful advantages to using phrases composed of stereotyped narrative text rather than constructing sentences directly from the UMLS vocabulary. First, the output text is more readable and closer to clinician's writing than methods of constructing sentences directly from data elements. Using phrases is more adaptable to the clinicians who are insistent about wording their notes a specific way. It can also effectively handle certain language subtleties such as the expression of confidence in a finding. Second, the use of phrases is potentially much faster for the clinician than constructing each phrase from the UMLS vocabulary. Of course, this increased speed is not gratuitous, and comes at the expense of time others must spend codifying the phrases. However, unlike coding each phrase from the UMLS vocabulary, the codification of phrases needs to occur only once. Third, the application is adaptable to almost any medical record system with an attached database that presents textual data in a window, and MEDPhrase can be used as a front end to many existing databases. Finally, the conceptual links to the phrases is independent of the language of the phrases, and foreign languages are just as easily supported as native languages.

There are disadvantages to using phrases composed of stereotyped narrative text rather than constructing sentences directly from the UMLS vocabulary. First, the use of modifiers within a phrase is not well supported. Separate phrases could be constructed, each with a single modifier replaced, but the number of phrases needed to express each instance increases with the number of modifiers in a sentence. Second, the phrases are not self-coding and the links to the controlled vocabulary need to be maintained when new phrases are added to the phrase libraries. The systemic use of the stereotyped phrases is a way of transferring the burden of codification from the clinician to other personnel. Future studies will need to center about the cost-efficiency of these alternate methods of coding the medical record.

Acknowledgments

The authors thank Mary Morgan for performing the queries of phrase occurrences in the COSTAR database. They also thank Dr. Mitchell Feldman for enlightening discussions and ongoing advice.

This work was supported in part by Training Grant NLM LM 07092 and in part by NLM research grant LM05854 and AHCPR grant HS06575. An equipment grant was provided by Hewlett Packard.

References

1. Cimino JJ, Barnett GO. The Physicians Workstation: Recording a Physical Examination Using a Controlled Vocabulary. Stead WW, ed. *Proceedings of the Eleventh Annual Symposium on Computer Applications in Medical Care,* pages 287-291, 1987.
2. Poon AD, Fagan LM, Shortliffe EH. The PEN-Ivory Project: Exploring User-interface Design for the Selection of Items from Large Controlled Vocabularies in Medicine. *JAMIA*, 3:168-183, 1996.
3. Sager N, Friedman C, Lyman MS. *Medical Language Processing, Computer Management of Narrative Data.* Addison-Wesley, 1987.
4. Huag PJ, Koehler S, Lau LM, et al. Experience with a Mixed Semantic/Syntactiv Parser. In Gardner RM, ed. *Proceedings of the Nineteenth Annual Symposium on Computer Applications in Medical Care*, pages 284-289, 1995.
5. Friedman C., Johnson SB, Forman B, Starren J. Architectural Requirements for a Multipurpose Natural Language Processor in the Clinical Environment. In Gardner RM, ed. *Proceedings of the Nineteenth Annual Symposium on Computer Applications in Medical Care*, pages 347-352, 1995.
6. DeFriece, RJ. Design Considerations for Intelligent Data Entry. In Gardner RM, ed. *Proceedings of the Nineteenth Annual Symposium on Computer Applications in Medical Care*, pages 91-96, 1995.
7. Somberg BL. A Comparison of Rule-based Positionally Constant Arrangements of Computer Menu Items. *Proceedings of Human Factors in Computing Systems and Graphics Interface*, pages 255-260, 1987.